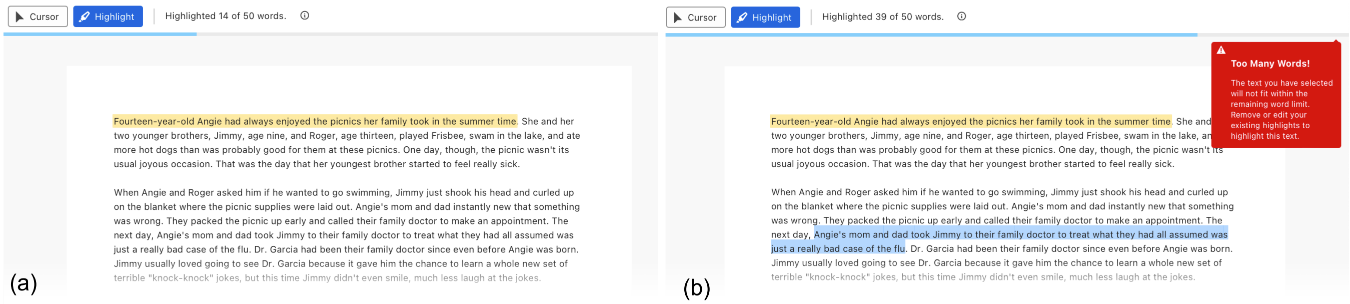
Constrained Highlighting in a Document Reader can Improve Reading Comprehension

Nikhita Joshi
nvjoshi@uwaterloo.ca
Cheriton School of Computer Science
University of Waterloo
Canada

Daniel Vogel
dvogel@uwaterloo.ca
Cheriton School of Computer Science
University of Waterloo
Canada



Fourteen-year-old Angie had always enjoyed the picnics her family took in the summer time. She and her two younger brothers, Jimmy, age nine, and Roger, age thirteen, played Frisbee, swam in the lake, and ate more hot dogs than was probably good for them at these picnics. One day, though, the picnic wasn't its usual joyous occasion. That was the day that her youngest brother started to feel really sick.

When Angie and Roger asked him if he wanted to go swimming, Jimmy just shook his head and curled up on the blanket where the picnic supplies were laid out. Angie's mom and dad instantly new that something was wrong. They packed the picnic up early and called their family doctor to make an appointment. The next day, Angie's mom and dad took Jimmy to their family doctor to treat what they had all assumed was just a really bad case of the flu. Dr. Garcia had been their family doctor since even before Angie was born. Jimmy usually loved going to see Dr. Garcia because it gave him the chance to learn a whole new set of terrible "knock-knock" jokes, but this time Jimmy didn't even smile, much less laugh at the jokes.

The text you have selected will not fit within the remaining word limit. Remove or edit your existing highlights to highlight this text.

Fourteen-year-old Angie had always enjoyed the picnics her family took in the summer time. She and two younger brothers, Jimmy, age nine, and Roger, age thirteen, played Frisbee, swam in the lake, an more hot dogs than was probably good for them at these picnics. One day, though, the picnic wasn't its usual joyous occasion. That was the day that her youngest brother started to feel really sick.

When Angie and Roger asked him if he wanted to go swimming, Jimmy just shook his head and curled up on the blanket where the picnic supplies were laid out. Angie's mom and dad instantly new that something was wrong. They packed the picnic up early and called their family doctor to make an appointment. The next day, Angie's mom and dad took Jimmy to their family doctor to treat what they had all assumed was just a really bad case of the flu. Dr. Garcia had been their family doctor since even before Angie was born. Jimmy usually loved going to see Dr. Garcia because it gave him the chance to learn a whole new set of terrible "knock-knock" jokes, but this time Jimmy didn't even smile, much less laugh at the jokes.

Figure 1: (a) Constrained highlighting interface. A count and progress bar showing how many words have been highlighted appear in the top toolbar; (b) if too many words are selected, an error message appears, and the new highlight is not created.

ABSTRACT

Highlighting text in a document is a common active reading strategy to remember information from documents. Learning theory suggests that for highlights to be effective, readers must be selective with what they choose to highlight. We investigate if an imposed user interface constraint limiting the number of highlighted words in a document reader can improve reading comprehension. A large-scale between-subjects experiment shows that constraining the number of words that can be highlighted leads to higher reading comprehension scores than highlighting nothing or highlighting an unlimited number of words. Our work empirically validates theories in psychology, which in turn enables several new research directions within HCI.

CCS CONCEPTS

• **Human-centered computing → Interaction tech**.

KEYWORDS

highlighting, reading, constraints, controlled experiments

ACM Reference Format:
Nikhita Joshi and Daniel Vogel. 2024. Constrained Highlighting in a Document Reader can Improve Reading Comprehension. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3613904.3642314

1 INTRODUCTION

Marking up existing text with underlines and highlights ("text-marking" [4]) is a common technique used by readers to remember information from documents [12]. Prior work in learning theory suggests this is caused by two main effects. First, marking up text visually isolates it from other text, making it more memorable [24]. Second, by considering whether some text is important and worth highlighting, readers think more about about it, resulting in better recollection [11, 31]. However, to reap these benefits, readers must be selective by only marking what is truly important [12, 24]. Yet many tend to over-mark text [4], such as when using text-marking to help concentration while reading [24]. This is problematic since it can create less visual separation between important and unimportant text, hindering recollection, and it can instill a false sense of comprehension [6, 31]. Effective text-marking strategies can be taught. For example, Leutner et al. [19] helped readers to reflect on their highlighting through self-regulation training, but this required lots of time and effort as readers had to follow a lengthy 90-minute training program consisting of almost 50 slides.

Constraints in design act as forcing functions on user behaviour [25, p. 141-145]. Seemingly arbitrary constraints applied to software can have positive effects, for example, by encouraging more participation on social media [16] or by promoting more focused knowledge-sharing [17, 22]. Imposing a constraint on text-marking is easy to do within document reader software, but whether this can lead to improved comprehension has not been examined. Enforcing hard limitations on how much text can be marked within a document reader should implicitly force readers to reflect on their markings and self-regulate. Specifically, the reader must reconsider whether each marking is truly important, and revise accordingly if the limit has been reached to regain the ability to mark new text.

This process forces the reader to play a more active role with text-marking and should lead to a highlighted document with adequate visual separation.

We conducted a large-scale between-subjects experiment (n=127) in which participants were assigned a level of text highlighting constraint: no highlights, up to 150 highlighted words, and unconstrained highlights. The results show that participants subjected to a highlight constraint performed better in a reading comprehension test taken 24 hours later. Our work contributes the first exploration of user interface constraints in the context of text-marking, and it is the first to show that constrained highlighting can improve reading comprehension scores without traditional self-regulation training. More broadly, our work validates theories in psychology, and applying this theory within HCI can lead to several research directions for the design of document reader interfaces.

2 BACKGROUND AND RELATED WORK

Our work is related to prior work on text-marking strategies in psychology and prior work that has imposed constraints or augmented highlights in text editors.

2.1 Benefits of Text-Marking

The reason why text-marking can be an effective active-reading strategy has been debated in psychology: is it because of the *act* or the end *result*? According to Levels of Processing theory [11], information is recalled for longer periods of time the deeper the information has been processed, which can be achieved through *"a greater degree of semantic or cognitive analysis,"* like making associations to prior knowledge and experiences. Yue et al. [31] gave students a passage to read and they were instructed to study for a test that took place one week later by either just reading the passage, or reading and highlighting the passage. Their results showed that students that highlighted less received higher scores than those who were considered heavy highlighters. Yue et al. speculated that being selective while highlighting required more mental effort to decide what to highlight, which led to higher scores.

The von Restorff effect [30] states that when presented with multiple items that are similar, either visually or semantically, items that differ are more likely to be remembered. This theory can explain why text-marking can help people recall information in documents. Nist and Hogrebe [24] gave students text passages with different types of information (i.e., important and unimportant details from the text) already underlined to examine whether text marks are beneficial for the resulting document they produce. Their results showed that when students are given passages with one type of information already underlined, they answered more questions about the underlined content correctly than students who received passages with another type of information highlighted.

Constraining text highlights could be beneficial for both the act and the result. Readers are encouraged to think more critically about what they highlight, which also results in a document with adequate visual separation.

2.2 Pitfalls of Text-Marking

For text-marking to be most effective, the reader should be able to distinguish between important and unimportant material [12]. Bell and Limber [4] explored the impact of reading skill on text-marking efficiency by examining the textbooks used by students in an introductory psychology course. Students with lower reading skills tended to over-highlight text, and highlighted more irrelevant information, which led to lower scores on the final exam than those with higher reading skills, who were more selective and focused when highlighting. Some readers may use text-markings for the wrong reasons, for example, as a concentration strategy while reading [24], which can lead to over-highlighting. Yue et al. [31] suggested that readers who do not know how to highlight effectively may feel a false sense of comprehension (i.e., "illusion of competence" [6]). When re-reading, these readers may skim over their highlights with little focus as they believe that the presence of a highlight means the content has already been encoded in memory.

Such pitfalls can be avoided by teaching readers how to highlight effectively. Using a computer training program, Leutner et al. [19] taught students how to highlight text using a five-step process, which included reflecting on their highlighting behaviours through self-regulation training. This encouraged students to monitor, self-evaluate, and make adjustments to their highlighting behaviours, and these students performed better in reading comprehension tests than those who just learned effective highlighting strategies. However, the training was lengthy, requiring readers to follow a slideshow consisting of almost 50 slides for roughly 90 minutes.

We show that constraining text highlights may encourage self-regulation by requiring the reader to monitor, self-evaluate, and adjust their highlights to adhere to the constraints imposed by document reader software.

2.3 Applications in HCI

To our knowledge, no prior work has investigated the impact of artificially constraining text-marking in a user interface. Some work has explored the positive impacts of short note-taking styles, like bullet journaling [7], on mindfulness and self-reflection [2, 29]. Biskjaer et al. [5] explored the effects of time constraints within a text editor to encourage more creative writing, and found that people wrote more when writing under a time constraint. Han et al.'s Textlets [14] turn text selections into interactive objects that can be manipulated and saved within a text editor to improve consistency when working under constraints imposed by technical documents. Although Textlets visually resemble text highlights, there were no limits on how many could be created within a document.

3 EXPERIMENT

The goal of this experiment is to understand the impact of constrained highlighting on reading comprehension scores. Participants read a short story and were asked to highlight text in preparation for an open-book reading comprehension test 24 hours later. This is a between-subjects study where each participant could either highlight nothing, highlight up to 150 words, or highlight an unlimited number of words.

3.1 Participants

We recruited participants through the Prolific crowdsourced experiment service.[1] Participants were restricted to Canada and the

[1] https://www.prolific.co

Table 1: Participant demographics and ways participants currently use document readers and highlight text inside document readers.

Gender		Age		Education		English Language Proficiency	
Men	62	18-24	5	Less than High School	2	Full Professional	6
Women	60	25-34	28	High School	15	Native or Bilingual	121
Non-binary	3	35-44	48	Some University (no credit)	18		
Unknown	2	45-54	20	Technical Degree	14		
		55-64	16	Bachelor's Degree	57		
		65-74	6	Master's Degree	12		
		75+	2	Beyond Bachelor's (e.g., MD, JD)	4		
		Unknown	2	Doctorate	5		

Document Reader Frequency		Highlight Frequency		Highlight Usage	
Daily	16	Daily	3	Remember Concepts	84
Weekly	44	Weekly	31	When Commenting	37
Monthly	29	Monthly	15	Concentration	43
Less than Monthly	28	Less than Monthly	50	Other	2
Never	10	Never	18		

United States and those who completed at least 2,500 tasks and with an approval rating greater than 98%. To identify fraudulent participant responses, we manually examined all open-ended responses for responses repeated across participants, or for very short, unrelated responses (e.g., "good" or "nice") [28]. No participants were omitted for this reason. Participants were instructed not to use any other study tools or aides while reading the document, like taking notes in a separate document or taking a screenshot of it. An open-ended response asked participants if they used any tools or aides to filter out those who did not follow the experiment instructions, which has been done in other crowdsourced experiments (e.g., [20]). In total, we filtered out 15 participants (11%) who described using other study tools or aides, who experienced technical difficulties with our user interface, and who did not attempt to answer any questions during the reading comprehension test, leaving 127 valid responses (Table 1). Participants received $15 in total. For each condition, participants who scored within the top 25% received a $3 bonus to provide a small incentive to do well on the test.

3.2 Task

Participants read one of ten short stories from easyCBM [1],[2] which is a system developed by the University of Oregon that provides teachers with benchmark assessments that were designed by researchers and school districts across the United States. The reading comprehension test in particular has been shown to predict student performance on state-wide exams and identify students who would likely not pass them [3]. All stories were roughly 1,500 words and were an 8th grade reading level, and have been used in prior HCI studies (e.g., [9]).

3.3 Apparatus

A Node.js and React application (Figure 1) served a web-based custom document reader with interfaces for reading and testing. The reading interface displayed the document at the centre of the screen. If the participant was able to highlight, at the top was a toolbar that allowed participants to change highlighting modes, which are common in existing document readers like Adobe Acrobat. Using the Cursor tool, participants could first select text from the document, and then press a black Highlight button. Using the Highlighter tool, any text selected automatically became highlighted. Text selections snapped to full words. If the participant could only highlight up to 150 words, the toolbar also displayed how many words had been highlighted, and a progress bar expanded and shrunk as words were highlighted or deleted. Copying text was disabled to prevent cheating.

The test interface displayed the same document with the participant's highlights with 20 multiple choice questions displayed on the right. The top toolbar displayed the number of questions answered, and a progress bar and countdown showed how much time remained for the test. The browser "find" feature was disabled on the document text to prevent cheating.

3.4 Identifying Experimental Properties

To identify levels of constrained highlighting, we first ran a pilot experiment without any highlighting constraints with 12 participants from Prolific. This was done to better understand how many words participants naturally highlight when reading the short stories we selected for the experiment. Overall, we observed that these participants highlighted 296 words on average (SD=201; Figure A.1). Based on these results, we initially selected 250, 150, and 50 word highlight limits, which corresponds to slightly below the mean followed by decreasing 100 word intervals.

With levels of constraint identified, we then ran a pilot to understand how much time to allocate to the reading comprehension test. We ran the experiment with a 10 minute time limit for the test with 59 participants (10 to 13 per condition). We found they spent 7.7 minutes on average to complete the test (SD=2.0) with a 15.1 average score (SD=3.1; Figure A.2). As the stories were short, participants were likely able to re-read them during the test, leading to higher scores and little differentiation between conditions. These results indicated a 5 minute time limit was reasonable to

[2] We received explicit permission from easyCBM to use the short stories and reading comprehension tests in our experiment.

increase test pressure and encourage participants to rely more on their memory and highlights, instead of re-reading the story.

As between-subjects experiments have less statistical power, it was not practical to run three experimental conditions alongside two baseline conditions at a large scale. As such, we ran another pilot with 98 participants (16 to 25 per condition) to identify which word limit was most promising. A shorter time limit proved to be successful at differentiating the conditions, and our results suggested that the 150 word limit may lead to higher scores when working under a 5 minute time limit for the test (Figure A.3). In the main experiment, we constrain highlights to 150 words, which corresponds to roughly 10% of the document word count. We include data from these participants in the main results.

3.5 Procedure

Participants received a link to the document reader web application through the Prolific system. The task was restricted to desktop and laptop devices. They entered basic demographic information and read instructions, then they were presented with the reading interface where they read the short story and highlighted content of interest if permitted in their condition. There was no time limit during the reading stage of the experiment. Once they finished reading, they answered 7 short questions about their experience using the reading interface.

After 24 hours, the participant could access the test interface, which displayed the short story marked up with their highlights along with 20 multiple choice questions. If the participant answered all questions within the 5 minute time limit, they could press a button to finish the test early. Otherwise, the test automatically ended after 5 minutes. Participants were not allowed to pause the timer during the test, and they were told to finish the test in one sitting prior to beginning. They answered 8 short questions about their experience completing the test.

3.6 Design

We opted for a between-subjects design over a within-subjects design to keep the experiment shorter for each participant and to prevent order effects across conditions (e.g., implicitly learning to highlight less, or learning the types of questions that may be asked in subsequent conditions). There is one primary independent variable, HIGHLIGHTS, with 3 levels: NONE (n=43), CONSTRAINED (i.e., up to 150 words; n=42), and UNCONSTRAINED (n=42). All participants read one of 10 documents, using one HIGHLIGHTS condition. Both were randomly assigned.

The primary measures computed from logs were:

- *Reading Comprehension*, the number of questions the participant answered correctly during the reading comprehension test (0-20 range).
- *Words per Highlight*, the number of words within a single highlight.
- *Total Words Highlighted*, the total number of words highlighted, counting only the highlights that the participant did not delete.
- *Number of Highlights*, the final count of highlights.
- *Duration*, the time taken (in minutes) to read or highlight the document.
- *Number of Deletions*, the number of times the participant deleted a highlight from the document.
- *Limit Reached*, an indicator variable for whether the word limit was exceeded while attempting to add a new highlight.

The post-reading questions had 6 measures from the NASA-TLX. The post-test questions had the same 6 measures and one additional question asking participants how frequently they referred back to the story and their highlights (all 1-7 scale). The values for *Performance* were reversed (i.e., 8 - x) to align valence and numeric scores. The post-reading and post-test questions both included a single open-ended question asking about the participant's experience.

4 RESULTS

Where applicable, we use a Kruskal-Wallis test and Mann-Whitney U tests with Holm's corrections for multiple comparisons. Error bars in charts are 95% confidence intervals (bootstrapped with 10,000 re-samples).

4.1 Reading Comprehension

Overall, we observe that constraining highlights to 150 words can improve reading comprehension scores (Figure 2). A significant main effect of *Reading Comprehension* ($\chi^2_{2,N=127} = 15.7$, $p < .001$, $\eta^2 = .11$) and post hoc tests revealed that CONSTRAINED (M=14.3, SD=3.4) led to higher scores than both NONE (M=10.5, SD=4.8; $p < .001$) and UNCONSTRAINED (M=12.1, SD=4.2; $p < .05$). Standard deviations and individual scores show that the spread of data for CONSTRAINED was tighter (IQR=13 to 17) than both NONE (IQR=7 to 14) and UNCONSTRAINED (IQR=10 to 15.75), suggesting that scores were more consistent when working under a CONSTRAINED HIGHLIGHTS constraint.

4.2 Highlighting Experience

To better understand why a 150 word highlight constraint improved *Reading Comprehension*, we grouped open-ended responses from the reading portion of the experiment for participants in the CONSTRAINED condition (n=42). The groupings were done by the first author as the data was straightforward.

Eighteen participants (43%) noted that the word limit affected their highlighting strategy. Specifically, sixteen (38%) indicated that the word limit encouraged them to highlight less and focus on the the most important points, with comments like: *"I kind of liked it*

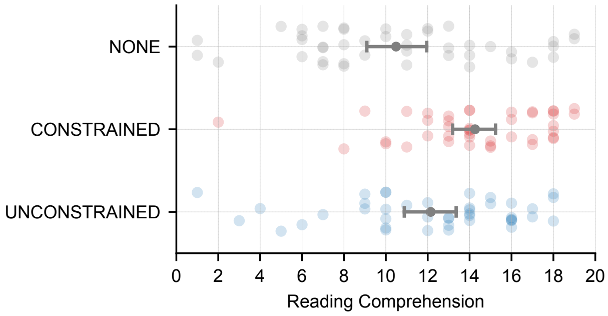


Figure 2: Individual and average *Reading Comprehension* by condition.

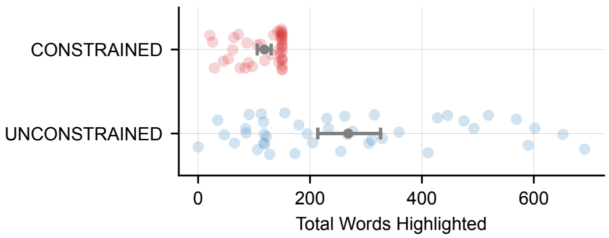
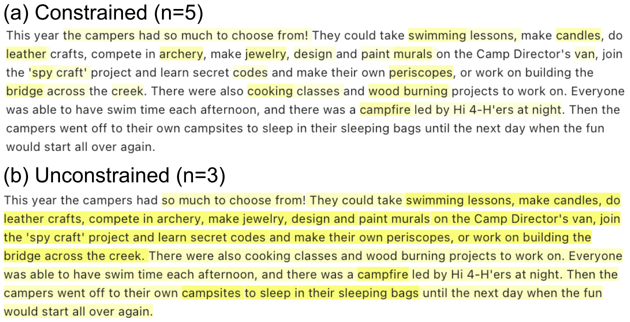


Figure 3: Individual and average *Total Words Highlighted* by condition.

because it forced me to highlight only the parts I thought were more important. In turn, this forced me to understand the story and main themes more" (P34).

4.2.1 Highlight Word Count. To corroborate these findings, we examined the *Total Words Highlighted* and *Number of Highlights* for the CONSTRAINED and UNCONSTRAINED conditions (Figure 3), and found that *Total Words Highlighted* was much lower in the CONSTRAINED condition (M=118.5, SD=41.9) than the UNCONSTRAINED condition (M=263.7, SD=186.5; $p < .001$). However, the *Number of Highlights* for the two conditions were similar (24.4 vs. 19.2 highlights), suggesting that each highlight contained fewer words. We examined the *Words per Highlight* to confirm this and found that when CONSTRAINED, highlights were an average of 4.8 words, much lower than 13.7 words in UNCONSTRAINED ($p < .001$).

4.2.2 Highlighted Content. To get a sense of the types of words participants highlighted, we examined heat maps of the raw highlights, where common highlights between participants appeared more opaque (Figure 4). As we had two highlighting conditions and participants could highlight one of ten stories, the number of participants who highlighted the same story for a single condition is low (5 to 10), so we discuss common themes across all stories. Participants highlighted a wide range of text, especially for the UNCONSTRAINED condition, so we filtered the heat maps to only show text where a majority of the participants highlighted the same thing (i.e., opacity ≥ 0.5) and compared similar types of highlights across conditions. One story had no participants for the CONSTRAINED condition, so we only consider nine stories.



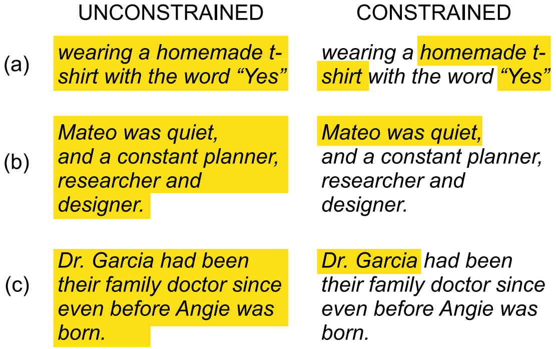
(a) Constrained (n=5)

This year the campers had so much to choose from! They could take swimming lessons, make candles, do leather crafts, compete in archery, make jewelry, design and paint murals on the Camp Director's van, join the 'spy craft' project and learn secret codes and make their own periscopes, or work on building the bridge across the creek. There were also cooking classes and wood burning projects to work on. Everyone was able to have swim time each afternoon, and there was a campfire led by Hi 4-H'ers at night. Then the campers went off to their own campsites to sleep in their sleeping bags until the next day when the fun would start all over again.

(b) Unconstrained (n=3)

This year the campers had so much to choose from! They could take swimming lessons, make candles, do leather crafts, compete in archery, make jewelry, design and paint murals on the Camp Director's van, join the 'spy craft' project and learn secret codes and make their own periscopes, or work on building the bridge across the creek. There were also cooking classes and wood burning projects to work on. Everyone was able to have swim time each afternoon, and there was a campfire led by Hi 4-H'ers at night. Then the campers went off to their own campsites to sleep in their sleeping bags until the next day when the fun would start all over again.

Figure 4: Example of highlighted content that was highlighted when (a) CONSTRAINED and (b) UNCONSTRAINED (opacity is normalized across participants).



	UNCONSTRAINED	CONSTRAINED
(a)	*wearing a homemade t-shirt with the word "Yes"*	*wearing a homemade t-shirt with the word "Yes"*
(b)	*Mateo was quiet, and a constant planner, researcher and designer.*	*Mateo was quiet, and a constant planner, researcher and designer.*
(c)	*Dr. Garcia had been their family doctor since even before Angie was born.*	*Dr. Garcia had been their family doctor since even before Angie was born.*

Figure 5: Example differences between similar highlights when UNCONSTRAINED and CONSTRAINED.

The most common differences between similar highlights for the two conditions were removing filler words when CONSTRAINED (7 stories; 78%); in contrast, this only occurred for two stories in the UNCONSTRAINED condition. This often involved separating longer highlights by article words or prepositions (Figure 5a). When additional adjectives were used to describe the same noun, the first adjective was typically highlighted while others were ignored (Figure 5b). When a person or place was described using a few sentences, participants highlighted the concept itself without the definition (Figure 5c). Although less information was highlighted when CONSTRAINED, some participants noted that each highlight effectively created a kind of bookmark to find additional details, for example, *"it was pretty handy in order to know where to look in the text for specific segments of the story [...] Due to the limit on how much I could highlight, I only really used it for that"* (P33). Some online resources at universities include highlighting tips like *"highlight key words and phrases instead of full sentences"* [26], so it appears as though participants in the CONSTRAINED were encouraged to highlight in this way.

4.2.3 Reaching the Limit. We anticipated that participants would delete more highlights in the CONSTRAINED condition once the limit was reached. Overall, it had twice as many deletions compared to UNCONSTRAINED (117 vs. 59). However, participants only reached the word limit 46 times (3% of all highlighting activities). This suggests that participants were highlighting few words from the onset rather than shortening retroactively, supported by comments like: *"the amount of highlighted words was something that I had to constantly keep track of. I predicted that I would run out of highlights available unless I used them carefully"* (P31).

This reluctance to delete highlights is further supported by the distribution of highlight locations for all valid highlights (Figure 6). Participants in the CONSTRAINED condition tended to highlight more at the beginning of the document, while those in the UNCONSTRAINED condition highlighted more consistently throughout the entire document. By focusing their highlights on text earlier in the document, participants typically ran out of words halfway through their reading and highlighting session (Figure 7), which led to more invalid highlighting attempts for text later in the document (Figure 8). One participant noted that running out of words at important moments of the story, which typically occurred halfway through

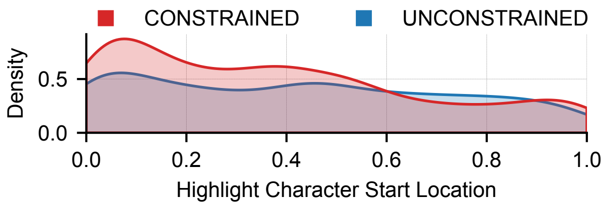


Figure 6: Distribution of highlight start locations by condition (normalized across documents).

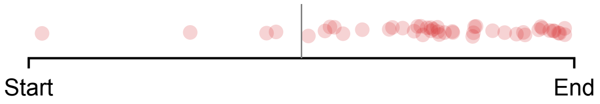


Figure 7: Timeline showing when participants in the CONSTRAINED condition hit the 150 word limit (normalized across participants).

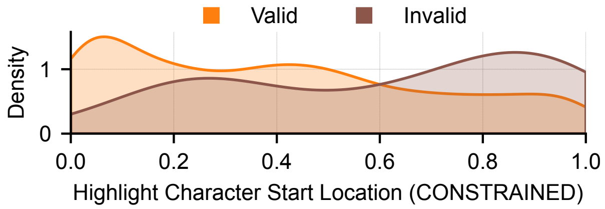


Figure 8: Distribution of valid and invalid highlighting attempts for the CONSTRAINED condition (normalized across documents).

the documents we selected, was frustrating: *"it was slightly frustrating that I ran out of words right when I got to the climax of the story"* (P75).

4.2.4 Duration. Although participants in the CONSTRAINED condition had to adopt new strategies and think more about highlighting under a word limit, they took roughly the same amount of time (M=12.3 minutes; SD=8.7) as those in the UNCONSTRAINED condition (M=11.2 minutes; SD=9.9; Figure 9). We anticipated that NONE would be faster than both CONSTRAINED and UNCONSTRAINED, but no significant differences were observed (M=8.9 minutes; SD=4.8).

4.3 Subjective Feedback

Overall, all conditions were rated similarly for all metrics post-reading (Figure 10) and post-test (Figure 11). Typically, average scores were below a "neutral" score of 4, with the exception of *Effort* for CONSTRAINED post-reading, all conditions for *Mental Demand*, *Temporal Demand*, and *Effort* post-test, and *Performance* for NONE post-test. Although eleven participants from the CONSTRAINED condition (26%) had indicated feelings of frustration or increased mental demand in open-ended responses after reading and highlighting the story, this did not seem to impact scores.

4.3.1 Document Types. Seven participants (17%) in the CONSTRAINED condition said that highlighting was not necessary for a short story. One participant even noted feeling so engaged with the text that they forgot to highlight: *"it helped me remember some important events, however, I now noticed that [when] reading the last parts of the study, I was so engrossed in it, I didn't highlight that much"* (P7). Two participants noted that their highlighting would be different for other types of documents, for example: *"I was not concerned about the limit. It was not the type of fact-rich text that typically would be highlighted. For example, historical or medical type texts are ones that I would expect to highlight"* (P6).

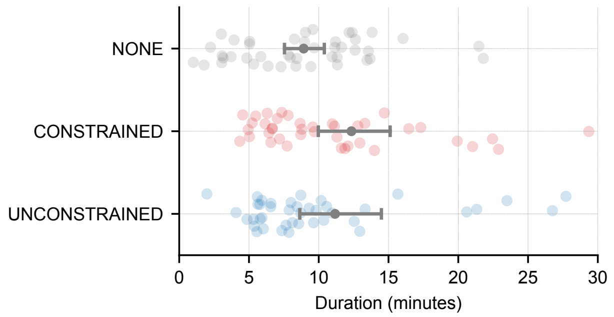


Figure 9: Reading *Duration* by condition. Note that 3 points with values greater than 30 minutes are not shown to improve visibility of the confidence intervals.

5 DISCUSSION

To summarize, our results show that constraining highlights can improve reading comprehension scores. Reading comprehension scores increased by 2.12 points (11%) when compared to having an unconstrained ability to highlight. Participants noted that having a word limit encouraged them to highlight only the most important points, and their highlights were in fact shorter and focused on highlighting key words like nouns, which is recommended by some university learning centres (e.g., [26]). This change in strategy did not increase reading time, nor did it increase mental demand, effort, or frustration when compared to an unconstrained ability to highlight. We discuss related research directions our work opens up for the broader HCI community and the limitations of our work.

5.1 Research Directions for HCI

A text highlight constraint is a very simple concept that could be integrated into existing document readers like Adobe Acrobat and macOS Preview. However, there remains open questions and design decisions for HCI researchers.

5.1.1 Identifying Constraints for Different Documents and Task Environments. During our pilot studies, we identified 150 words (roughly 10% of the document word count) as a promising constraint for the types of short stories and the type of task we selected: a reading comprehension test with additional pressure from a short time limit. However, as suggested by participants and our own pilot testing (see Section 3.4 and Appendix A for details), this is not a universal solution as different documents and tasks will require different levels of constraint. One possibility is to allow users to set their own levels of constraint for individual documents, but a more ambitious goal would be to analyse document characteristics to

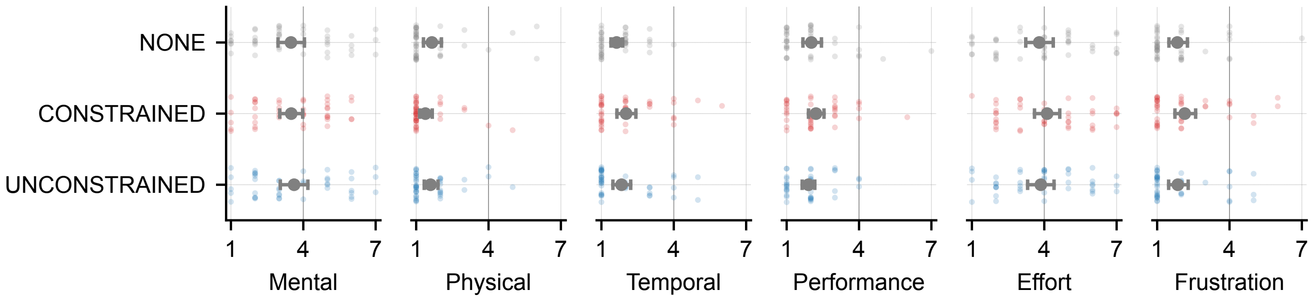


Figure 10: Questionnaire scores by condition after the reading portion of the experiment. Lower scores are better.

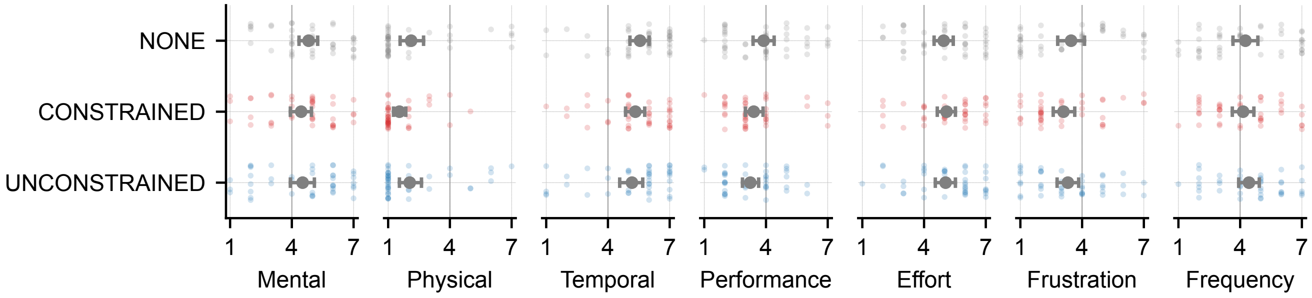


Figure 11: Questionnaire scores by condition after the test portion of the experiment. Lower scores are better.

calculate an optimal level of highlight constraint. Constraints could be imposed relative to the text structure, for example, allowing only 5 highlights per section, or allowing more highlights for certain types of sections, like the results section of an academic paper. This might encourage readers to highlight more consistently throughout a document, which was something our participants seemed to struggle with.

Studying for a test has a clear objective, but text may be highlighted for a variety of reasons, many of which are more exploratory or ambiguous [8]. For example, knowledge workers may frequently switch between broadly capturing text to gather information and narrowing or filtering text to create meaningful insights [15, 27]. Similarly, when collaborating with others, a reader will likely highlight more text initially before filtering their personal highlights to share with others [23]. A text highlight constraint should adapt to suit the nature of these tasks, allowing for more words to be highlighted for exploratory or ambiguous tasks, and less words for narrowing or filtering tasks.

Text highlight constraints could help readers learn. For example, in a classroom setting, teachers could set levels of constraints to encourage better study habits among students. The level of constraint could even act as training, where a document reader first learns current highlighting behaviours, then gradually imposes a more constrained word limit over time to help readers to develop better highlighting strategies.

5.1.2 Integration with Existing Features. Recall 37 participants (29%) reported previous highlighting experience linked to adding a text comment (Table 1). Imposing word limits on other types of document annotations like comments may have similar benefits, but it is unclear how text highlight constraints should be combined with text comment constraints since they are often simultaneous. One option is to have separate word limits for highlights and comments, allowing for a separation of concerns. Another option is to have a combined word limit (Figure 12a). Both options increase user effort: either keeping track of multiple limits or allocating words across these two features.

5.1.3 Interaction Techniques. Highlight constraints could be augmented or enhanced by the way they are created in the interface. For example, while highlighting, users could indicate a level of importance for each highlight by layering multiple strokes over the same text or by applying more pressure when using a pen. The least important highlights could automatically disappear once the limit has been reached (Figure 12b). This would prevent the user from having to manually delete highlights retroactively.

In this work, we explored "hard" constraints, since the interface strictly enforced the word limit, but "soft" constraints that merely act as suggestions that are not enforced could be used instead. For example, delays akin to those incorporated into marking menus [18] could deter readers from over-highlighting, much like they can improve learning of expert commands and keyboard shortcuts [13, 20]. Other ways to make highlighting slightly more difficult or effortful once the limit is reached, like slightly obfuscating the text in a frost-brushing interface [10] (Figure 12c), could also encourage users to stay within the recommended highlight constraint. For exploration tasks, the fuzzy boundaries of intentionally uncertain highlights [8] could shrink after the reader revisits a document to encourage them to filter important information.

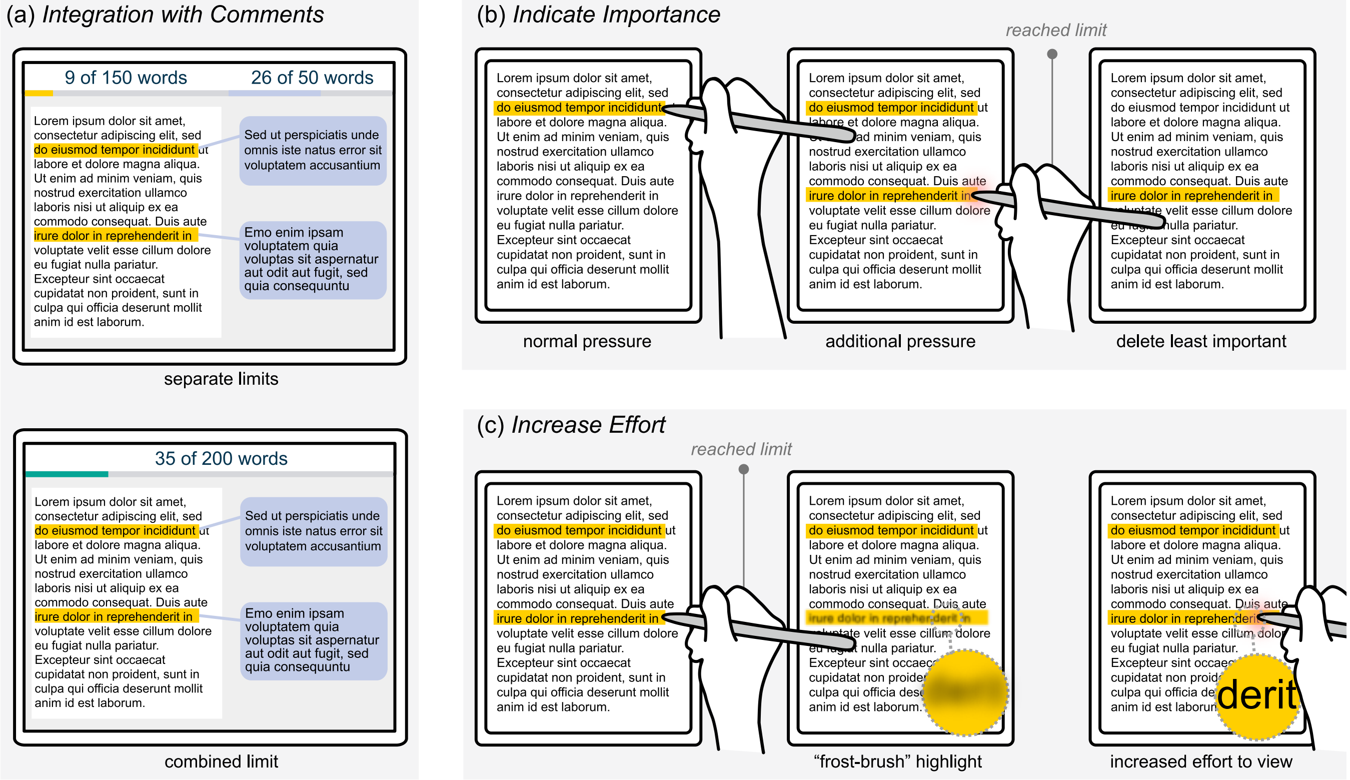

(a) Integration with Comments
9 of 150 words
26 of 50 words
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.
Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium
Emo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntu
separate limits
35 of 200 words
combined limit
(b) Indicate Importance
reached limit
normal pressure
additional pressure
delete least important
(c) Increase Effort
reached limit
derit
"frost-brush" highlight
increased effort to view

Figure 12: Interaction techniques to augment or enhance constrained highlighting.

5.2 Limitations

We tested multiple stories for improved external validity, but this reduced the number of participants per story for each condition. This made it difficult to analyse the raw highlight text and differences between conditions in greater depth. Repeating this study with a single story or greatly increasing the number of participants would allow us to learn what text is commonly highlighted between participants and give better insights into how people change strategies when highlighting under a word limit. Our results suggest that constrained highlighting could teach readers effective highlighting techniques, but this could be further validated by comparing it to an unconstrained highlighting condition in which participants learn effective highlighting techniques before reading a document.

Although we tried our best to mimic what it would be like to study for a test, our experimental setup may be lacking in ecological validity. Specifically, Lonka et al. [21] note that study strategies used during an experiment may be different than those used when studying for an actual exam. The documents we selected allowed for high internal validity and roughly correspond to something a student may face in an English course. However, there are other types of documents where text-marking is arguably even more useful, possibly with additional text-marking tendencies (e.g., non-fiction articles). A longitudinal study of text highlight constraints within a real educational setting would further validate and extend our findings.

Prior work suggests that people with lower reading abilities struggle to identify key concepts to highlight [4]. Although we conducted a study with a large population with a diverse educational background, we did not formally measure reading ability. This would have required additional time-consuming tasks, like the Nelson-Denny reading test, which would greatly increase the duration and fatigue of our experiment. It is likely that constrained highlighting would be more difficult and mentally demanding for this population, but perhaps once mastered, they may experience the greatest improvements in reading comprehension.

6 CONCLUSION

Using a large-scale, between-subjects experiment, we show that a text highlight constraint can improve reading comprehension scores when compared to not highlighting anything or unlimited highlighting. Our work validates theories in psychology, which state that being more selective when highlighting text improves recollection. At its core, the idea of constraining text highlights is incredibly simple. However, we believe that incorporating it into existing document reader software is an "easy win" that can help people become better learners by forcing them to be more selective and intentional with their highlights without the need for lengthy and time-consuming self-regulation training. Furthermore, it can open up several opportunities that would be of interest to the broader HCI community. In the context of text highlights, "less is more," and we hope our work will inspire new features and interactions within document reader software that are designed around constraints.

ACKNOWLEDGMENTS

We thank Damien Masson for his help debugging our web application. This work was made possible by NSERC Discovery Grant 2018-05187 and Canada Foundation for Innovation Infrastructure Fund 33151 "Facility for Fully Interactive Physio-digital Spaces."

REFERENCES

[1] Julie Alonzo, Gerald Tindal, Kirt Ulmer, and Aaron Glasgow. 2006. easyCBM® online progress monitoring assessment system. (2006). Eugene, OR: University of Oregon, Behavioral Research and Teaching.

[2] Amid Ayobi, Tobias Sonne, Paul Marshall, and Anna L. Cox. 2018. Flexible and Mindful Self-Tracking: Design Implications from Paper Bullet Journals. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal, QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173602

[3] Doris Luft Baker, Gina Biancarosa, Bitnara Jasmine Park, Tracy Bousselot, Jean-Louise Smith, Scott K Baker, Edward J Kame'enui, Julie Alonzo, and Gerald Tindal. 2015. Validity of CBM measures of oral reading fluency and reading comprehension on high-stakes reading assessments in Grades 7 and 8. *Reading and Writing* 28 (2015), 57–104. https://doi.org/10.1007/S11145-014-9505-4

[4] Kenneth E Bell and John E Limber. 2009. Reading skill, textbook marking, and course performance. *Literacy Research and Instruction* 49, 1 (2009), 56–67. https://doi.org/10.1080/19388070802695879

[5] Michael Mose Biskjaer, Jonas Frich, Lindsay MacDonald Vermeulen, Christian Remy, and Peter Dalsgaard. 2019. How Time Constraints in a Creativity Support Tool Affect the Creative Writing Experience. In *Proceedings of the 31st European Conference on Cognitive Ergonomics* (Belfast, United Kingdom) *(ECCE '19)*. Association for Computing Machinery, New York, NY, USA, 100–107. https://doi.org/10.1145/3335082.3335084

[6] Robert A Bjork. 1999. Assessing our own competence: Heuristics and illusions. *Attention and performance XVII: Cognitive regulation of performance: Interaction of theory and application* (1999), 435–459.

[7] Ryder Carroll. 2018. *The Bullet Journal Method: Track the Past, Order the Present, Design the Future.* Penguin.

[8] Joseph Chee Chang, Nathan Hahn, and Aniket Kittur. 2016. Supporting Mobile Sensemaking Through Intentionally Uncertain Highlighting. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (Tokyo, Japan) *(UIST '16)*. Association for Computing Machinery, New York, NY, USA, 61–68. https://doi.org/10.1145/2984511.2984538

[9] Xiuge Chen, Namrata Srivastava, Rajiv Jain, Jennifer Healey, and Tilman Dingler. 2023. Characteristics of Deep and Skim Reading on Smartphones vs. Desktop: A Comparative Study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, 14 pages. https://doi.org/10.1145/3544548.3581174

[10] Andy Cockburn, Per Ola Kristensson, Jason Alexander, and Shumin Zhai. 2007. Hard lessons: effort-inducing interfaces benefit spatial learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '07)*. Association for Computing Machinery, New York, NY, USA, 1571–1580. https://doi.org/10.1145/1240624.1240863

[11] Fergus I.M. Craik and Robert S. Lockhart. 1972. Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior* 11, 6 (1972), 671–684. https://doi.org/10.1016/S0022-5371(72)80001-X

[12] Robert L Fowler and Anne S Barker. 1974. Effectiveness of highlighting for retention of text material. *Journal of Applied Psychology* 59, 3 (1974), 358–364. https://doi.org/10.1037/h0036750

[13] Tovi Grossman, Pierre Dragicevic, and Ravin Balakrishnan. 2007. Strategies for accelerating on-line learning of hotkeys. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '07)*. Association for Computing Machinery, New York, NY, USA, 1591–1600. https://doi.org/10.1145/1240624.1240865

[14] Han L. Han, Miguel A. Renom, Wendy E. Mackay, and Michel Beaudouin-Lafon. 2020. Textlets: Supporting Constraints and Consistency in Text Documents. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376804

[15] Nanna Inie, Jonas Frich, and Peter Dalsgaard. 2022. How Researchers Manage Ideas. In *Proceedings of the 14th Conference on Creativity and Cognition* (Venice, Italy) *(C&C '22)*. Association for Computing Machinery, New York, NY, USA, 83–96. https://doi.org/10.1145/3527927.3532813

[16] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. 2007. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis* (San Jose, California, USA) *(WebKDD/SNA-KDD '07)*. Association for Computing Machinery, New York, NY, USA, 56–65. https://doi.org/10.1145/1348549.1348556

[17] Nikhita Joshi, Justin Matejka, Fraser Anderson, Tovi Grossman, and George Fitzmaurice. 2020. MicroMentor: Peer-to-Peer Software Help Sessions in Three Minutes or Less. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI 2020)*. Association for Computing Machinery, New York, NY, USA, 1–13.

[18] Gordon Kurtenbach and William Buxton. 1994. User learning and performance with marking menus. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, Massachusetts, USA) *(CHI '94)*. Association for Computing Machinery, New York, NY, USA, 258–264. https://doi.org/10.1145/191666.191759

[19] Detlev Leutner, Claudia Leopold, and Viola den Elzen-Rump. 2007. Self-regulated learning with a text-highlighting strategy. *Zeitschrift für Psychologie/Journal of Psychology* 215, 3 (2007), 174–182. https://doi.org/10.1027/0044-3409.215.3.174

[20] Blaine Lewis, Greg d'Eon, Andy Cockburn, and Daniel Vogel. 2020. KeyMap: Improving Keyboard Shortcut Vocabulary Using Norman's Mapping. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3313831.3376483

[21] Kirsti Lonka, Sari Lindblom-YlÄnne, and Sini Maury. 1994. The effect of study strategies on learning from text. *Learning and Instruction* 4, 3 (1994), 253–271. https://doi.org/10.1016/0959-4752(94)90026-4

[22] Xing Lu and Zhicong Lu. 2019. Fifteen Seconds of Fame: A Qualitative Study of Douyin, A Short Video Sharing Mobile Application in China. In *Social Computing and Social Media. Design, Human Behavior and Analytics: 11th International Conference, SCSM 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26-31, 2019, Proceedings, Part I* (Orlando, FL, USA). Springer-Verlag, Berlin, Heidelberg, 233–244. https://doi.org/10.1007/978-3-030-21902-4_17

[23] Catherine C. Marshall and A. J. Bernheim Brush. 2004. Exploring the relationship between personal and public annotations. In *Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries* (Tuscon, AZ, USA) *(JCDL '04)*. Association for Computing Machinery, New York, NY, USA, 349–357. https://doi.org/10.1145/996350.996432

[24] Sherrie L Nist and Mark C Hogrebe. 1987. The role of underlining and annotating in remembering textual information. *Reading Research and Instruction* 27, 1 (1987), 12–25. https://doi.org/10.1080/19388078709557922

[25] Don Norman. 2013. *The Design of Everyday Things: Revised and Expanded Edition.* Basic books.

[26] University of North Carolina at Chapel Hill. 2016. Highlighting – Learning Center — learningcenter.unc.edu. https://learningcenter.unc.edu/tips-and-tools/using-highlighters/. Accessed 30-08-2023.

[27] Peter Pirolli and Stuart Card. 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of International Conference on Intelligence Analysis*, Vol. 5. McLean, VA, USA, 2–4.

[28] Timothy Ryan. 2020. Fraudulent responses on Amazon Mechanical Turk: A Fresh Cautionary Tale. https://timryan.web.unc.edu/2020/12/22/fraudulent-responses-on-amazon-mechanical-turk-a-fresh-cautionary-tale/

[29] Jakob Tholander and Maria Normark. 2020. Crafting Personal Information - Resistance, Imperfection, and Self-Creation in Bullet Journaling. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376410

[30] William P Wallace. 1965. Review of the historical, empirical, and theoretical status of the von Restorff phenomenon. *Psychological Bulletin* 63, 6 (1965), 410–424. https://doi.org/10.1037/h0022001

[31] Carole L Yue, Benjamin C Storm, Nate Kornell, and Elizabeth Ligon Bjork. 2015. Highlighting and its relation to distributed study and students' metacognitive beliefs. *Educational Psychology Review* 27, 1 (2015), 69–78. https://doi.org/10.1007/s10648-014-9277-z

A APPENDIX

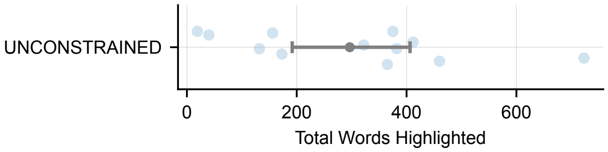


Figure A.1: Individual and average *Total Words Highlighted* for the first pilot (n=12) to identify word limits to test in subsequent pilot studies.

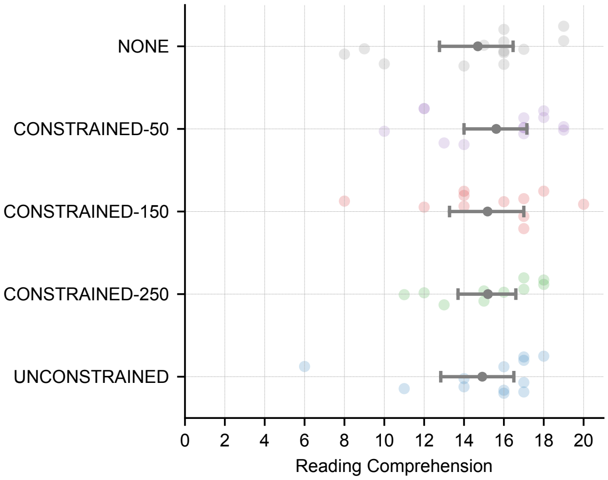


Figure A.2: Individual and average *Reading Comprehension* by condition for the second pilot (n=59) featuring a 10 minute time limit for the reading comprehension test.

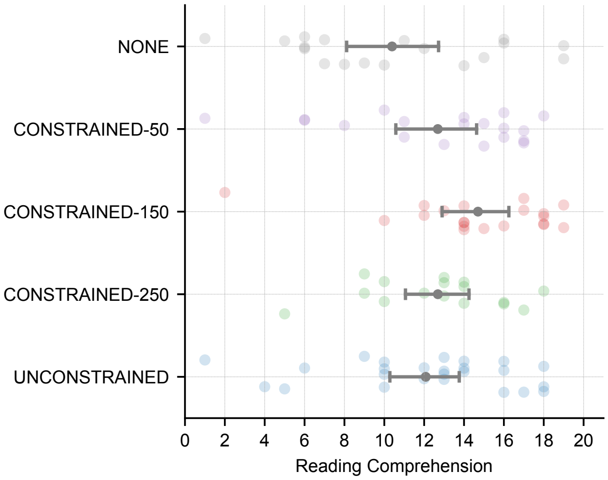


Figure A.3: Individual and average *Reading Comprehension* by condition for the third pilot (n=98) featuring a 5 minute time limit for the reading comprehension test.